Prozódiai információ használata az automatikus beszédfelismerésben; mondat modalitás felismerése

Vicsi Klára, Szaszák György és Németh Zsolt

Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium, 1111 Budapest, Sztoczek u. 2.
{vicsi, szaszak}@tmit.bme.hu

Kivonat: A mai, statisztikai elvi alapokra épülő folyamatos gépi beszédfelismerők kimenetén szóláncok sorozata jelenik meg, tehát a beszédfelismerés több szintű feldolgozási folyamatából a szószintig jutott el a mai beszédfelismerési technológia. Robusztus beszédfelismerés eléréséhez azonban további – például szemantikai – szintek bevonása szükséges.

A beszéd szupraszegmentális (prozódiai) paramétereinek bevonásával egy olyan prozódiai felismerőt hoztunk létre, amely a mondatok és tagmondatok fajtáit, azaz modalitását, illetve a mondatok határait ismeri föl, és ezzel hozzájárulhat a szemantikai szintű nyelvi felismerés biztosabb döntéseihez. Ez az ún. modalitás felismerő statisztikai elven működik, a mondatok, tagmondatok intonációs struktúráját leíró Rejtett Markov modellekből, és egy igen egyszerű, a mondatok kapcsolódására vonatkozó modellből épül fel.

A felismerő tesztelési eredményei azt mutatták, hogy azoknál a modalitás típusoknál, amelyekre a statisztikai betanításhoz elegendő minta állt rendelkezésre, a helyesen felismert modalitás aránya 75 és 95% között változott az adott mondat modalitásától függően.

1 Bevezetés

A beszédfelismerési folyamatnak számos szintje létezik: akusztikai, fonetikai-fonológiai, szintaktikai, szemantikai, pragmatikai szint (Ainsworth 1976). Ezek közül a szintek közül minél többet tudunk a gépi beszédfelismerési folyamatba bevonni, annál biztosabb lesz a felismerés.

A gépi beszédfelismerésnél akusztikai szinten működik az akusztikai előfeldolgozó egység, amely a beszédjel elemzését, a lényegkiemelést, tömörítést végzi el. Kimenetén jelennek meg az időkeretenkénti lényegi paraméterek (jellemző vektorok), amelyek szegmentális akusztikai szintű előfeldolgozásnál jellemzően 10 ms időkeretekben, 25-50 ms időablakban mért színképi paraméterek, leggyakrabban MFC (Mel-frekvencia kepsztrális) együtthatók (ld. 1. ábra 'a' feldolgozási ága) (Young 2005). A hagyományos beszédfelimerési folyamatban a következő szinten, a fonetikai-fonológiai szinten történik a beszéd szegmentális feldolgozása, vagyis az akusztikai szinten kapott lényegi paraméterek segítségével végezzük el a beszéd-

hangok modelljeinek a megalkotását. Felismeréskor az akusztikai szinten kapott jellemző vektorsorozatot hasonlítjuk össze a beszédhangok modelljeivel. Ez az a feldolgozási szint, ahol a kimeneten fonémasorozatot kapunk. Amennyiben szintaktikai szintű nyelvtant is bekapcsolunk a felismerésbe – például a legelterjedtebben használt statisztikai alapú N-gram[1] nyelvi modelleket –, akkor a kimeneten szósorozatot kapunk (Becchetti–Ricotti 1999), amint az az 1. ábra 'a' ágában látható. A kereskedelemben manapság elérhető beszédfelismerők így működnek.

2 Célkitűzés

A Beszédakusztikai Laboratóriumban olyan vizsgálatokat végeztünk, amelyekben arra kerestük a választ, hogy az akusztikai előfeldolgozással hogyan lehet hozzájárulni a magasabb feldolgozási szintek, a szintaktikai, valamint szemantikai szintű nyelvi feldolgozás eredményesebbé tételéhez. Az akusztikai előfeldolgozásra ekkor már nem célszerű a fent említett szegmentális tartományban végzett lényegkiemelés jellemző vektorait használni. Más, szupraszegmentális (prozódiai) jellemzőkön alapuló lényegkiemelésre van szükség, amely tükrözi bizonyos beszédtartalmak megkülönböztetését, az értelmi tagolást és akár az érzelmeket is. Ennek megfelelően a beszéd fizikai paramétereit a szupraszegmentális tartományban, jellemzően durvább frekvencia- és időfelbontásban célszerű vizsgálnunk, mint amikor a szegmentális jellemzés volt a cél. A szupraszegmentális jellemzők vizsgálatánál figyelembe kell vennünk néhány tényt, amelyek nehezítik e tartományban a jellemző vektorok kinyerését. Ezek közül néhányat alább közlünk.

– A szupraszegmentális paraméterek jelentős mértékben variálódnak a beszédstílus, a beszélő, a tartalom, a környezet, stb. függvényében.

– A nyelv rétegződését a nyelv sok esetben a szupraszegmentális információval is érzékelteti. Az üzenet nyelvi rétegződési szintjei és a szupraszegmentális információ között azonban nagymértékű az egymásra hatás, jellemző a szintek közötti kapcsolatok bonyolultsága (Langlais, 1993). Egy adott szinten jellemző szupraszegmentális jellemzőket nehéz kinyerni, mivel a felette lévő szint erősen befolyásolja az alsóbb szintek alakulását. Közismert például, hogy a szóhangsúlyokat a mondathangsúly erősen befolyásolja.

– A szupraszegmentális jellemzésre használt fizikai paraméterek közül az alapfrekvencia pontos mérésének nehézségei, illetve rendszerbe illesztésük közismert.

Amikor a szupraszegmentális paramétereket (illetve a belőlük előállított jellemző vektorokat) használjuk a beszédfelismerés segítésére, ezt több szinten tehetjük meg, amint ez az 1. ábrán a 'b' és a 'c' ágban látható. A magasabb nyelvi szintek felé haladva, lépésről lépésre járulunk hozzá a beszédfelismerés biztonságosabbá tételéhez. Szintaktikai szinten, a hagyományos szólánc kimenetű felismerők teljesítményéhez

[1] A beszédfelismerésben nyelvi modell alatt egy adott nyelv lehetséges szókapcsolatainak leírását értjük. Statisztikai nyelvi modell esetén az „N-gram" elnevezés arra utal, hogy a nyelvi modell adott, N darab szóból felépülő szósorozatok előfordulási valószínűségeit tárolja paraméterként.

képest lényegesen javíthatjuk a felismerés biztonságát kötött hangsúlyozású nyelveknél (pl. magyar vagy finn nyelvekre) a szóhatárok automatikus bejelölésével. Ez a

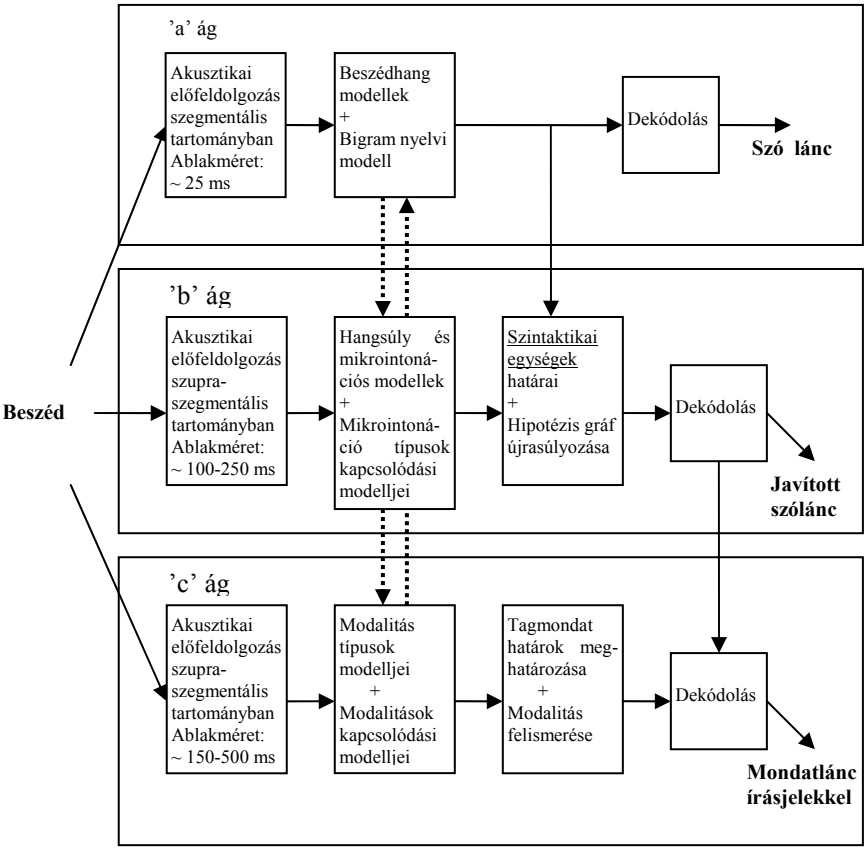


Fig. 1: A kibővített több szintű beszédfelismerő tömbvázlata (az ábra részletes magyarázatát lásd a fenti szövegben

szintaktikai szintű feldolgozás az 1. ábra 'b' feldolgozási ágában követhető végig. A módszer lényege, hogy olyan kötött hangsúlyozású nyelveknél ahol a szóhangsúly az első szótagon van, szupraszegmentális paraméterek segítségével mikrointonációs egységeket tudunk automatikusan a beszédfolyamba bejelölni, amelyek mutatják a szóhatárt. A magyar nyelvre beszédpercepciós vizsgálatok is megerősítik e módszer helyességét, hiszen az emberi beszédfeldolgozó rendszerünk is használja ezeket a szupraszegmentális információkat a szóhatárok megtalálásához (Tóth 2007). A szóhatárok automatikus bejelölésének módszeréről, a felismerésbe történt beépítés hatékonyságáról már korábban beszámoltunk (Vicsi–Szaszák 2004, 2005).
Az 1. ábra 'c' feldolgozási ágában szemantikai szintű feldolgozás látható. Jelen cikkünk tárgyát éppen ez a szupraszegmentális paraméterekre épülő szemantikai szintű

feldolgozás képezi, amely során az esetleges tagmondatok határainak lokalizációját és a mondatok modalitásának felismerését hajtjuk végre. A modalitás felismerésnél az dönthető el, hogy például a feltételezetten kimondott „*alma van a fa alatt*” szósor állítás, kérdés vagy felkiáltás formájában hangzott el. A mondat és tagmondat határok lokalizációja révén támpontot is kapunk a tekintetben, hogy melyek a szósorozatban a (tag)mondatok kezdő- és végidőpontjai.

3 Szematnikai szintű modalitás felimserő elvi alapjai, a fejlesztés módszere

A mondat-, ill. tagmondatfajták felismerését statisztikai elven, a Rejtett Markov Modell (HMM) módszer alapján végeztük el, amelyre a HTK fejlesztői rendszert (Young 2005) használtuk. A felismerő betanításához a modalitásfajták szerint feldolgozott beszédadatbázist használtuk. Megjegyezzük, hogy a tagmondatokat is el kellett különítenünk modalitás szempontjából aszerint, hogy az a mondat, amelybe beágyazódnak, milyen modalitású. A továbbiakban tehát – némi pongyolasággal – tagmondatok modalitásáról (ill. fajtáiról) is írunk majd, ez alatt természetesen mindig az értendő, hogy az adott tagmondat milyen modalitású mondat része. Az egyes (tag)mondattípusokra HMM modelleket építettünk fel, amelyek segítségével a tagmondatfajtákat felismertük. A felismeréshez felhasználtuk a tagmondatok egymás utáni sorrendjét figyelembe vevő szöveg szintű prozódiai modellt is, mely tulajdonképpen egy egyszerű nyelvi szabálymodell. Alapvető feladatnak tekintettük a különböző HMM tagmondatfajták modelljeinek optimális beállítását.

3.1 A betanító anyag elkészítése

A BABEL (Vicsi et al. 1998) és az MRBA (Vicsi et al. 2004) beszédadatbázisainkból kigyűjtöttük a különböző mondatfajtákhoz tartozó mondatokat. Összesen 10 alapvető mondat, ill. tagmondatfajtát (típust, modalitást) különböztettünk meg (vö. Olaszy 2002) az 1. Táblázat szerint.A kiválasztott hangfájlokat lehallgatás és az alapfrekvencia, illetve energiaszint mérése alapján tagmondatok szerint szegmentáltuk és címkéztük a (tag)mondatok „modalitása” szerint, vagyis a hanganyagba bejelöltük a mondatok, tagmondatok határait, valamint a tagmondatok típusainak (modalitásainak) megfelelő szimbólumokat. Szegmentálásra, címkézésre a 2. ábrán mutatunk példát. A 2. ábra első sorában a hanganyag hullámformája, a második sorban az alapfrekvencia és az intenzitás görbéi láthatók. A harmadik sorban van a kézzel elvégzett szegmentálás és címkézés. A mondat és tagmondat típusokon, illetve ezek határain kívül a tagmondatok és mondatok közötti szünetrészt külön bejelöltük (’U’ szimbólummal). A szünetrész bejelölésére a tagmondatok között a kb. 400 ms-nál, míg a mondathatároknál a kb. 500 ms-nál nagyobb szüneteknél került sor – lásd például a 2. ábrán a két kijelentő (’S’-sel jelölt mondat) közti kiemelt részt. Ezen értékeknél kisebb szüneteknél általában csak határt jelöltünk, és a határ feléhez tettük be az „elválasztást” mint a 2. ábrán az ’E’-vel jelölt eldöntendő kérdés, és a ’T’-vel jelölt tagmondat között.

1. Táblazat: A szegmentálás és címkézés statisztikája

Egyszerű mondat modalitás és összetett mondat modalitása tagmondatonként[2]	**Jelölés (cimke)**	**Összes előfordulás (db)**
Kijelentő mondat, Kijelentést záró tagmondat	S	445
Kijelentő tagmondat, a záró tagmondat nélkül	T	287
Kiegészítendő kérdés	K	40
Kiegészítendő kérdés tagmondata	KT	13
Eldöntendő kérdés	E	35
Felszólító és felkiáltó mondatok	FF	52
Felszólító és felkiáltó mondatok tagmondata	FFT	24
Óhajtó mondat	O	2
Felsorolás	F	41
Semleges	N	125
Összesen:		**1029**

Az ilyen módon feldolgozott adatbázissal végeztük el a prozódiai felismerőnk betanítását. Mivel az óhajtó mondatra csak 2 mintánk volt az adatbázisban, ezt a tanításnál kihagytuk. Így 9 tagmondat és egy szünet HMM modellt hoztunk létre. A szegmentálás és címkézés statisztikáját az 1. táblázat mutatja.

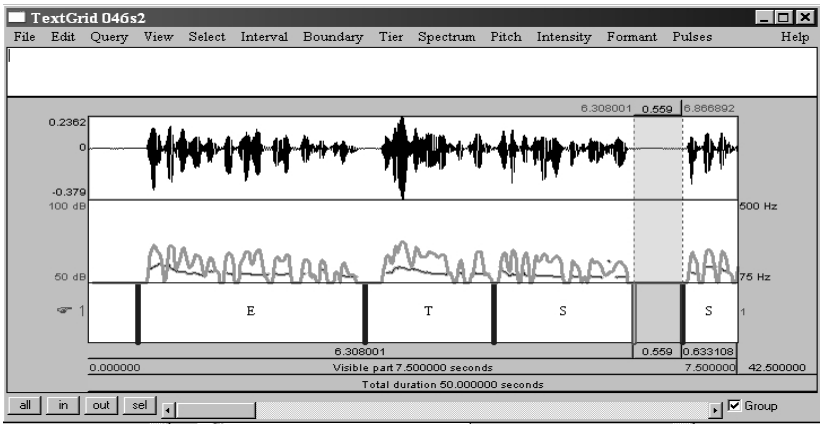


Fig. 2: Szegmentálás és címkézés a Praat programban.

[2] Tagmondat modalitáson jelen cikkünkben, a beszédtechnológiai kezelhetőség érdekében, most azt értjük, hogy egy adott tagmondat milyen modalitású összetett mondat része.

Egyszerű és összetett mondatokat vegyesen használtunk fel, és a táblázat statisztikájából látható, hogy több mint 1000 címkét helyeztünk el szegmentálás közben.

Az eredeti energia- (e_i) és alapfrekvencia-értékeket (f_{oi}) 25 ms időablakban, 10 ms-os időkeretenként mértük. Az alapfrekvencia értékét gördülő átlagos magnitudókülönbség-függvény (Short-time Average Magnitude Difference Function – AMDF) segítségével határoztuk meg (Gordos et al 1983).

Az előfeldolgozás utolsó lépéseként a mért energia és alapfrekvencia értékeket különböző időablakban átlagoljuk. Ezek az 5, 10, 20, 26, 30, 36, 40 és 50 keretszámok, szorozva a 10 ms keretidővel. A bizonyos intervallumban átlagolt érték lesz a középső elem értéke. A különbözőképpen betanított modellek közül teszteléskor választjuk ki az optimálisat.

Az alapfrekvencia értékeinek a feldolgozásakor oktávszűrést hajtunk végre, mivel a felhangszerkezetben téveszthet az alapfrekvenciát detektáló algoritmus: oktávot ugorhat.

Az e_i és f_{oi} értékek mellett három intervallum nagyság alapján 3-3 első és második deriváltat számítottunk ki mind az alapfrekvenciához, mind az intenzitáshoz. Az intervallum nagyságát a deriváltak számítására szolgáló regressziós képlet szerint vettük figyelembe:

$$d_t = \frac{\sum_{i=1}^{T} i\left(c_{t+i} - c_{t-i}\right)}{2\sum_{i=1}^{T} i^2}. \qquad (1)$$

A (1) képlet a t időpillanathoz számít derivált értéket, a c_t a t-hez tartózó együtthatót jelenti. A T az intervallum nagyságát jelentő változó, amelynek értéke 10, 20 és 40 lesz.

Így keletkezik az összesen 14 elemű jellemzővektor:
$V_{jell}=\{f_{oi}, e_i, df_{oi}^{10}, d^2f_{oi}^{10}, df_{oi}^{20}, d^2f_{oi}^{20}, df_{oi}^{40}, d^2f_{oi}^{40}, de_i^{10}, d^2e_i^{10}, de_i^{20}, d^2e_i^{20}, d^2e_i^{40}, d^2e_i^{40}\}$.
A d, d^2 az első és a második deriváltat, míg a deriváltak utáni indexben lévő szám az intervallum nagyságát jelenti.

3.2 Betanítás a különböző paraméterekkel

A feldolgozott beszédadatbázist két részre bontottuk: az egyik résszel a betanítást, míg a másikkal a tesztelést végeztük. A mondatok nagyrészt véletlenszerűen lettek kiválasztva, de arra odafigyeltünk, hogy minden felismerendő címke szerepeljen mind a betanított, mind a tesztelésre szánt anyagban.

A betanítás során az adatbázis hangfájljaiból az előfeldolgozással nyert szupraszegmentális jellemző vektorokat, valamint az adatbázis szegmentálási és címkézési adatait használjuk fel a prozódiai modellek felépítéséhez.

3.3 Tesztelés

A szemantikai szintű prozódiai felismerő tesztelése két fő folyamatból áll: a felismerésből és az összehasonlítás utáni értékelésből. A folyamatot a 3. ábra szemlélteti. A szupraszegmentális előfeldolgozás után, a betanítás során kialakított tagmondattípus modelleket használtuk a felismeréshez, valamint a korábban már említett, mondatok kapcsolódását leíró szabályokat. Ebben olyan nyelvtani szabályokat írtunk elő a modalitás felismeréséhez, hogy az a hétköznapi, folyamatos beszédben gyakran, kevés kivétellel előforduló eseteket maradéktalanul lefedje. A szabályok azt adják meg, hogy milyen (tag)mondat milyen (tag)mondatot követhet, és milyen (tag)mondatot nem, illetve milyen (tag)mondatok ismétlődhetnek, stb. Lényegében analóg szerepe van a beszédfelismerésnél használt nyelvi modellével, de statisztikai adatok hiányában, illetve a jóval kevesebb lehetőség miatt szabályokat adtunk meg.

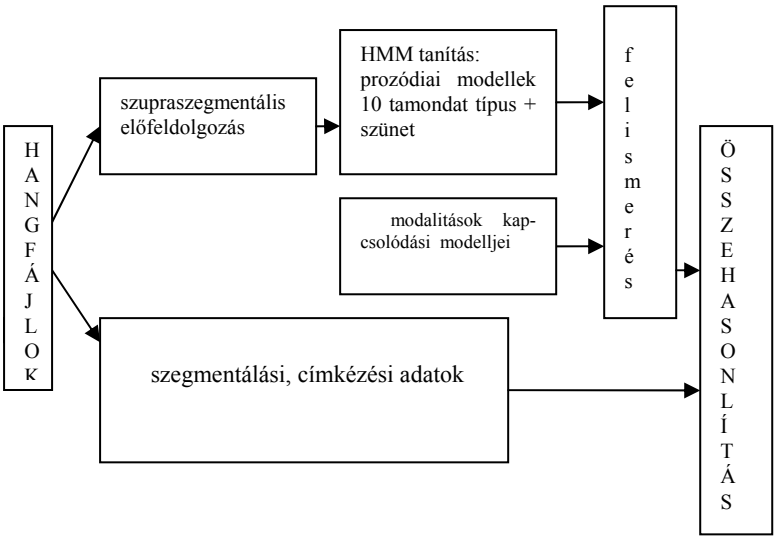


Fig. 3: A tesztelés folyamatábrája

A felismerés jóságát a helyes felismerés arányával *(Corr)* és a pontossággal *(Acc*) adjuk meg. A modalitás szerint helyesen felismert tagmondatok aránya:

$$Corr = \frac{H}{N} \cdot 100\% \tag{2}$$

A pontosság számítása:

$$Acc = \frac{H - I}{N} \cdot 100\% \tag{3}$$

ahol *H* a modalitás szerint helyesen felismert tagmondatok, *I* a beszúrások[3] és *N* az összes tagmondat száma.

A teszteléseket először az 1. táblázat szerinti 10 különböző címke betanításával és felismerésével kezdtük. Az eredmények feldolgozása során hamar kiderült, hogy egyes modalitás típusokból nincs elegendő számú minta a betanításhoz, valamint az energia- és alapfrekvencia-menet „hasonlósága" miatt egyébként is célszerű csoportosítást végezni az alábbiak szerint:

- A felsorolások (’F’), illetve a felkiáltó és felszólító mondatok tagmondataiból (’FFT’) is arányaiban kevés minta van, továbbá intenzitás- és alapfrekvencia-szerkezetükben is hasonlítanak. Ezeket gyakran felsorolásnak (’F’), vagy tagmondatnak (’T’) detektálja a felismerő. Továbbá mindegyikhez a vessző írásjel tartozik, ezért mindkét csoport összevonható a kijelentő mondat tagmondatával: ’F’, ’FFT’ → ’T’.
- A kiegészítendő kérdést tartalmazó mondatok tagmondatainak (’KT’) szerkezete nagy hasonlóságot mutat az egyetlen tagmondatból álló kiegészítendő kérdésével (’K’). A mondatok értelmezése szempontjából továbbá nem jelent különbséget, ha a *„Hová menne, és mit csinálna akkor?"* mondatot két kérdésként: *„Hová menne? És mit csinálna akkor?"* ismeri fel a modell. Ezért ezek összevonhatóak: ’KT’ → ’K’.

Így végül a csoportosítással (összevonással) 6 tagmondatmodellt, és egy szünet-modellt tanítottunk be, és használtunk a felismeréshez.

A további teszteléskor a szupraszegmentális jellemző vektorok átlagolási intervalluma (lásd a „Betanító anyag előkészítése" c. pontban), valamint a HMM (tag)mondatmodellek állapotainak száma függvényében vizsgáltuk a felismerés jóságát. Arra kerestük a választ, hogy az energia és alapfrekvencia jellegű jellemzők milyen időfelbontása szükséges ahhoz, hogy a különböző tagmondattípusokat optimálisan tudjuk felismerni.

Kérdés továbbá a HMM tagmondatmodellek állapotainak optimális száma. Nyilvánvalóan több állapotra van szükség, mint a fonémamodelleknél használt 3 állapot, de hogy ezen modelleknél hány állapotot kell felvennünk az optimális felismeréshez, azt a teszteléssel döntöttük el.

A két tényezőt együttesen vizsgáltuk, rögzített ($\log P_{ins}=0$) szóbeszúrási valószínűséggel[4] végeztük el a modalitás felismerést (ld. 4. ábra). Az eredményeket a 2. táblázat szemlélteti. A táblázat oszlopaiban találhatóak az átlagolási intervallum keretszámai. A sorokban a tagmondattípus modellekben beállított állapotok száma a változó értéke. A táblázat celláiban – százalékban kifejezve – találhatóak a helyes felismerés (*Corr*) eredményei.

[3] Beszúrásnak nevezik a beszédfelismerésben a valós tesztanyagban nem megjelenő, azonban a felismerő által feltételezett és így tévesen felismert elemet, mely esetünkben annak felel meg, hogy a modalitás felismerő egy további tagmondatokra nem bontható (tag)mondatot tévesen felbontott.

[4] A szóbeszúrási valószínűség a beszédfelismerők egy állítható paramétere, melynek kisebbre állításával – durva megfogalmazással – a felismerő mérsékli az adott hangmintára illesztett szimbólumok számát. Esetünkben a szóbeszúrási valószínűség a „(tag)mondat beszúrás" valószínűségének felel meg.

2. Táblázat A 7 különböző címke helyes felismerésére (*Corr*) %-osan

HMM Álla-potok száma		Időablak 10 ms-os keretenként							
		5	**10**	**20**	**26**	**30**	**36**	**40**	**50**
	5	-	-	60,24	59,76	60,00	58,31	58,31	58,80
	11	66,07	67,47	67,95	68,92	67,47	67,23	69,40	65,06
	15	-	66,99	66,51	66,27	67,47	66,99	64.58	66,47
	19	-	-	66,99	64,10	65,06	63,37	63,37	60,02

A legjobb eredményt 11 HMM állapot mellett kaptuk. Az időablak nem változtatja tendenciózusan az eredményeket 100 és 400 ms között. A legjobb átlagos modalitásfelismerés 69,4 % volt.

A tagmondattípusokra lebontott tévesztési mátrix a 11-es állapotszám és a 40 keretnyi átlagos intervallum mellett a 3. táblázatban látható.

A mátrix sorai jelentik azt, hogy mi volt az eredeti modalitás, az oszlopok jelentése pedig, hogy mit ismert fel a felismerő. Az utolsó, '*Ins'* feliratú sorban lévő tagmondat típusokat hamisan beszúrta a felismerő, és a '*Del'* feliratú oszlopban lévőket pedig törölte.

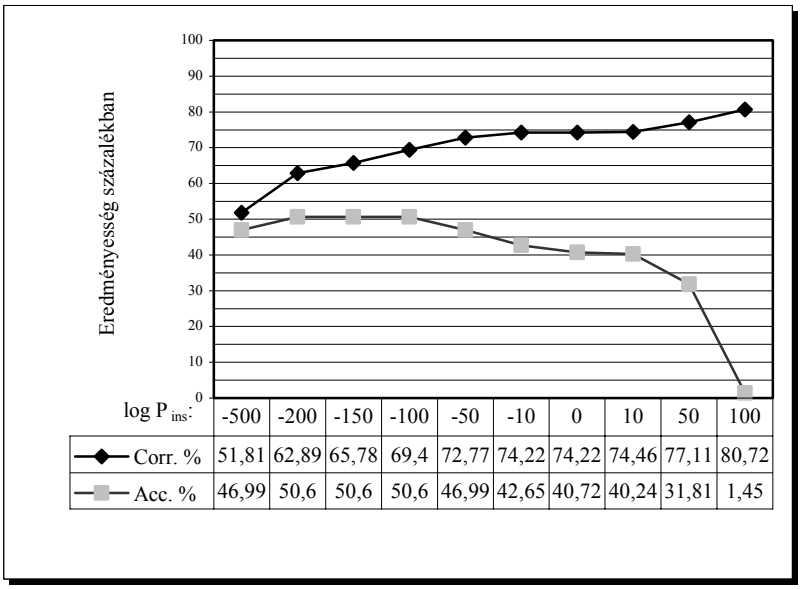
A tévesztési mátrix jobb oldalán látható, hogy bizonyos mondatfajtákra egészen elfogadható eredmények születtek: '**FF**' – 50%, '**T**' – 83,3%, '**S**' – 74,8%, és '**U**' – 96,0%. Ezek közül az első három érdemleges a modalitás-típusok felismerése, az utolsó pedig a mondathatárok detektálása szempontjából.

3. Táblázat. A tagmondattípusokra lebontott tévesztési mátrix a 11-es állapotszám és a 40 keretnyi átlagolási intervallum mellett, Corr=69.40 %, Acc=50.60 %

	S	**T**	**K**	**E**	**F**	**N**	**U**	**Del**	**[corr[%]]**
S	**83**	11	7	4	2	3	1	7	[74.8]
T	4	**70**	0	1	2	0	7	13	[83.3]
K	3	3	**4**	0	0	0	2	3	[33.3]
E	1	1	0	**2**	1	0	1	0	[33.3]
FF	0	2	1	0	**5**	0	2	4	[50.0]
N	3	4	0	0	2	**4**	2	14	[26.7]
U	0	5	0	0	0	0	**120**	11	[96.0]
Ins	6	32	3	2	4	4	27		

A (tag)mondat beszúrás valószínűségének optimalizálása is fontos szerepet játszik a felismerés hatékonyságában. Érdemes azonban odafigyelni arra is, hogy a helyesen felismert tagmondatok mellett a pontosság is fontos. A beszúrás valószínűségének növelésével a helyesen felismert modalitások mellett sok plusz címkét is elhelyezünk,

így például a mondathatárok meghatározása nagyon nehézkessé válik. A két eredményességi mutató változásait ezért együttesen figyeltük a tagmondatbeszúrás logaritmusának (a 4. ábrán $\log P_{ins}$) függvényében, amint azt a 4. ábra mutatja.



log P_{ins}:	-500	-200	-150	-100	-50	-10	0	10	50	100
Corr. %	51,81	62,89	65,78	69,4	72,77	74,22	74,22	74,46	77,11	80,72
Acc. %	46,99	50,6	50,6	50,6	46,99	42,65	40,72	40,24	31,81	1,45

Fig.4: Az eredményesség alakulása a (tag)mondatbeszúrás valószínűségéneklogarimusa ($\log P_{ins}$) függvényében

A helyesen felismert tagmondat modalitás típusok aránya a 100-as értéknél adódik a legmagasabbra. Az is megfigyelhető, hogy ekkor a pontosság százaléka nagyon alacsony (mindössze 1,45%), ami azt jelenti, hogy a felismerés tele van „felesleges" beszúrásokkal, és a mondathatárokat nem lehet helyesen felismerni. A pontosság három vizsgált értéknél haladja meg az 50%-ot (mindháromnál 50,6%), és ezek közül a helyesen felismert szavak aránya a -100-as értéknél a legmagasabb: 69,4%.

4 Értékelés

A fentiekben bemutatott szemantikai szintű modalitás felismerő nem túl nagy, és mondattípus eloszlásban is egyenetlen adatbázissal lett betanítva és tesztelve. Ennek ellenére a paraméterek optimális beállítása mellett a vártnál jobb eredmények adódtak. A legjobb felismerési eredményt akkor kaptuk, amikor az energia és az alapfrekvencia időfelbontása 100-400 ms közötti volt (átlagos olvasott beszédtempó mellett), a HMM tagmondattípus modellek állapotainak száma 11, és a mondatelem beszúrás valószínűségének a logaritmusa: -100 volt.

Ezen beállításokkal közel 70% a helyesen felismert címkék aránya, és a pontosság is több mint 50%-os értéket mutat, annak ellenére, hogy egy-egy tagmondatból több

száz darab csak az **S** és **T** mondatfajtáknál fordult elő. Az ’**S’** és ’**T’** típusok (kijelentő mondat és tagmondata) kb. 75%-os, illetve 83%-os eredménnyel detektálhatóak, továbbá az ’**FF’** mondatok helyes felismerése is eléri az 50%-ot annak ellenére, hogy betanításra, tesztelésre összesen csak 52 mondatunk volt. Továbbá a mondathatárok 96% pontosságú bejelölése szintén jó eredmény (ld. a.3. táblázat ’U’ feliratú sorában). A mondathatárok automatikus (’rec’) valamint kézi (’lab’) bejelölése összehasonlítására mutatunk példát az 5. ábrán.

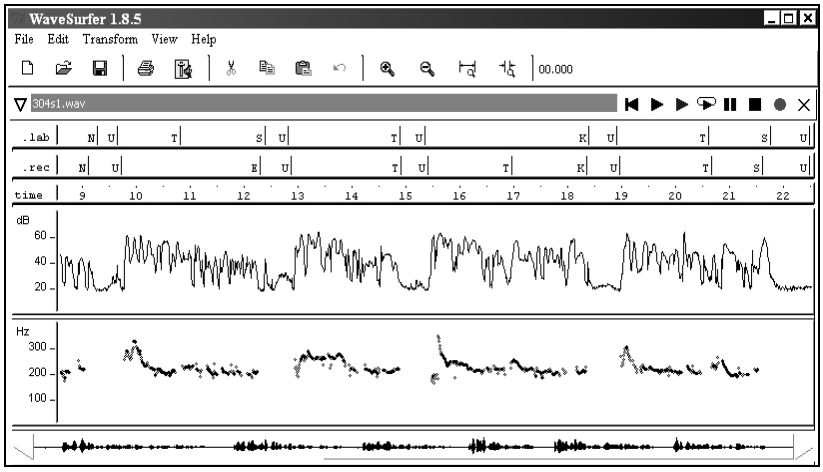


Fig.5. Tagmondathatárok kézi (’lab’) és automatikus (’rec’) bejelölése

Az eredményekből látható, hogy a nagyméretű adathalmazzal betanított tagmondatfajták (kijelentő mondatok és a kijelentő mondatok tagmondata) jó eredménnyel felismerhetőek, ezért célszerű a további mondatfajtákhoz is hasonló mennyiségű betanító és tesztelő anyag feldolgozása a jövőben.

A tagmondat alapú nyelvi modell nagyméretű adathalmazzal történő, statisztikai alapú kialakítása szintén jelentősen javíthatná a felismerés biztonságát.

Specifikusabbá lehet tenni a felismerőt, ha előre szerkesztett párbeszédekből felépült adatbázis betanításával építhetnénk fel a tagmondat modelleket, mert akkor az érzelmek (például a felkiáltó és felszólító mondatok prozódiai tulajdonságai) jobban kimutathatóak lennének, mint a most használt olvasott szöveg adatbázisokban végzett válogatás alapján.

A mondat és tagmondat típusok, valamint a mondathatárok felismerésének javítására a munkát tovább kell folytatni. Jelen cikkünkkel az volt a célunk, hogy bemutassuk, hogy érdemes ezen jellemzők vizsgálata, és bevonáshasznos az automatikus gépi beszédfelismerésbe.

Bibliográfia

1. Ainsworth, W.: Mechanisms of speech recognition. Pergamon Press. Oxford. (1976) 110-124.
2. Becchetti, C.-Ricotti, L. P: Speech Recognition,Theory and C++ implementation. Fondazione Ugo Bordoni, John Wiley. Rome .(1999)
3. Gordos, G.-Takács, Gy.: Digitális beszédfeldolgozás. Műszaki Könyvkiadó. Budapest. (1983) 239-240.
4. Gósy, M.: Fonetika, a beszéd tudománya. Osiris. Budapest. (2004) 182-243.
5. Langlais, P.-Méloni, H: Integration of a prosodic component in an automatic speech recognition system. 3rd European Conference on Speech Communication and Technology. Berlin. .(1993) 2007-2010.
6. Tóth, L.: Benchmarking Human Performance on the Acoustic and Linguistic Subtasks of ASR Systems. INTERSPEECH2007, Antverp. (2007) 382-385.
7. Olaszy, G.: A magyar kérdés dallamformáinak és intenzitásszerkezetének fonetikai vizsgálata. Beszédkutatás. Gósy Mária. MTA Nyelvtudományi Intézet. Budapest, (2002) 83-99.
8. Young, S., et al.: The HTK Book (for HTK Version 3.3). Cambridge University. Engineering Department, (2005)
9. Vicsi, K., Víg, A.: Az elsõ magyar nyelvû beszédadatbázis. Beszédkutatás. Gósy Mária. MTA Nyelvtudományi Intézete. Budapest. (1998) 163-177.
10. Vicsi Klára-Kocsor András-Teleki Csaba-Tóth László: *Beszéd adatbázis irodai számítógép-felhasználói környezetben*. II. Magyar Számítógépes Nyelvészet Konferencia. Szeged. Alexin Zoltán-Csendes Dóra. Szegedi Tudományegyetem. Informatikai Tanszékcsoport. (2004) 315-318.
11. Vicsi Klára-Szaszák György-Borostyán Gábor: *Folyamatos beszéd szó- és frázisszintű automatikus szegmentálása szupraszegmentális jegyek alapján*. II. Magyar Számítógépes Nyelvészet Konferencia. Szeged. Alexin Zoltán-Csendes Dóra. Szegedi Tudományegyetem Informatikai Tanszékcsoport. (2004) 319-326.
12. Vicsi Klára-Szaszák György: *Folyamatos beszéd szó- és frázisszintû automatikus szegmentálása szupraszegmentális jegyek alapján.* II. rész *Statisztikai eljárás, finn - magyar nyelvû összehasonlító vizsgálat.* Magyar Számítógépes Nyelvészeti Konferencia. Szeged. Alexin Zoltán-Csendes Dóra. Szegedi Tudományegyetem. Informatikai Tanszékcsoport. (2005) 360-370.